

Harmonic Analysis of Boolean Networks: Determinative Power and Perturbations

Reinhard Heckel*

Steffen Schober†

Martin Bossert†

September 6, 2011

Abstract

Consider a large Boolean network with a feed forward structure. Given a probability distribution for the inputs, can one find—possibly small—collections of input nodes that determine the states of most other nodes in the network? To identify these nodes, a notion that quantifies the *determinative power* of an input over states in the network is needed. We argue that the mutual information (MI) between a subset of the inputs $\mathbf{X} = \{X_1, \dots, X_n\}$ of node i and the function $f_i(\mathbf{X})$ associated with node i quantifies the determinative power of this subset of inputs over node i . To study the relation of determinative power to sensitivity to perturbations, we relate the MI to measures of perturbations, such as the *influence* of a variable, in terms of inequalities. The result shows that, maybe surprisingly, an input that has large influence does not necessarily have large determinative power. The main tool for the analysis is Fourier analysis of Boolean functions. Whether a function is sensitive to perturbations or not, and which are the determinative inputs, depends on which coefficients the Fourier spectrum is concentrated on. We also consider *unate* functions which play an important role in genetic regulatory networks. For those, a particular relation between the influence and MI is found. As an application of our methods, we analyze the large-scale regulatory network of *E. coli* numerically: We identify the most determinative nodes and show that a small set of those reduces the overall uncertainty of network states significantly. The network is also found to be tolerant to perturbations of its inputs, which can be seen from the Fourier spectrum of its functions.

1 Introduction

A Boolean network (BN) is a discrete dynamical system, which is for example used to study and model a variety of biochemical networks such as genetic regulatory networks. BNs have been introduced in the late 1960s by Kauffman [1, 2], who proposed to study random BNs as models of gene regulatory networks. Kauffman investigated their dynamical behavior and a phenomena called self-organization. Aside from its original purpose, BNs were also used to model (small-scale) genetic regulatory networks, e.g. in [3, 4, 5] it was demonstrated that BNs are capable of reproducing the underlying biological processes (i.e., the cell-cycle) well. Boolean models are also used to model large-scale networks, such as the *Escherichia coli* regulatory network [6] which is analyzed in Section 6. This network is, in contrast to Kauffman’s automata and the regulatory

*Department of Information Technology and Electrical Engineering, ETH Zurich

†Institute of Telecommunications and Applied Information Theory, University of Ulm

networks considered in [3, 4, 5], not an autonomous system, since the gene’s states are determined by external factors.

Concerning the analysis of BNs, measures of perturbations have received most attention. Whether a random Boolean network operates in the so called ordered or disordered regime is determined by whether a single perturbation, i.e., flipping the state of a node, is expected to spread or die out eventually. Kauffman [2] argues that biological networks must operate at the border of the ordered and disordered regime, hence they must be tolerant to perturbations to some extent. To our knowledge, a measure of determinative power of a node has neither been studied systematically nor related to measures of perturbations. Let us give an example, where a notion of determinative power is needed: Consider a feed-forward network where the states of the nodes are controlled by the states in the input layer. Now we ask, if a possibly small set of inputs suffice to determine most states, i.e., reduces the uncertainty about the network’s states significantly. For the *E. coli* regulatory network (see Section 6) this answers whether a small set of metabolites and other inputs determine most genes that account for *E. coli*’s metabolism. The setting in this paper is as follows. The state of each node in the network is viewed as an independent random variable with some distribution. This assumption applies e.g. for networks with a tree-like topology, and is a standard assumption when studying the effect of perturbations. For this setting, determinative power of nodes and perturbation-related measures are properties of single functions, hence the analysis of the BN reduces to the analysis of single functions. As the main tool, we use Fourier analysis of Boolean functions. Fourier analytic techniques were first applied to Boolean networks by Kesseli et al. [7, 8]. They used spectral techniques to derive results related to Derrida plots of random BNs and the convergence of trajectories in random BNs. In this paper, we discuss mutual information as a measure of determinative power. Ribeiro et al. [9] considered the pairwise mutual information in time series of random Boolean networks. The setup considered here is different from the one studied in [9], as in [9], the functions are chosen at random, whereas here the functions are given, but the argument is random. Furthermore, we are interested in a measure of decisive power of subsets of inputs over the function’s output, statistical dependencies, and connections to perturbations, whereas in [9] the pairwise mutual information in time series of random Boolean networks is studied.

Contributions. We argue that the mutual information between a set of nodes and the state of a node is a measure of the determinative power of this set of inputs. The argument is as follows: Mutual information is a quantity that measures the mutual dependence of random variables. If a set of inputs to a node and the state of this node are strongly mutually dependent, then this set can be viewed as having large determinative power over this node. To understand determinative power and mutual dependencies in Boolean networks better, we systematically study mutual information of a sets of inputs and the state of a node. We relate the mutual information to measures of perturbations, and prove that—maybe surprisingly—a set of inputs that is highly sensitive to perturbations, must not have determinative power at all. Conversely, an input that has determinative power, must be sensitive to perturbations to some extent. These results are proven by expressing a Boolean function f as a multilinear polynomial over the reals, i.e., considering the Fourier expansion of f . Then both measures of perturbations and of determinative power can be simply computed from the Fourier spectrum of a Boolean function. Specifically we argue that the concentration of weight in the Fourier domain on sets of inputs characterizes a function in terms of tolerance to perturbations and determines the determinative power of nodes. Furthermore we generalize a result given by

Xiao and Massey [10] that gives a necessary and sufficient condition of statistical independence of sets of inputs in terms of the Fourier coefficients. This result can e.g. be applied to decide for which classes of functions the algorithm presented in [11], that detects functional dependencies based on estimating mutual information, can in principle detect dependencies, or fails. Moreover we show, that for a *unate* function, any input and the function’s output are statistically dependent. For unate functions we also prove a direct relation between the mutual information and the influence of a variable. The class of unate functions includes all linear threshold functions and describes functional dependencies in gene regulatory networks well [12]. As an application of the theoretical results in this paper, we show that mutual information can be used to identify the determinative nodes in the large-scale model of the control network of *E. coli*’s metabolism [6].

Outline. The paper is organized as follows. Boolean networks and Fourier analysis of Boolean functions are reviewed in Section 2. In Section 3, influence and average sensitivity as measures of perturbations are reviewed, and their relation to the Fourier spectrum is discussed. In Section 4 we study the mutual information of sets of inputs and the function’s output. In Section 5, we discuss the class of unate functions. Section 6 contains an analysis of the large-scale *E. coli* regulatory network, using the tools and ideas developed in previous sections.

2 Preliminaries

In this section we review some standard facts about Boolean networks and Fourier analysis of Boolean functions and introduce notation.

2.1 Boolean Networks

A (synchronous) Boolean network (BN) can be viewed as a collection of N nodes with memory. The state of a node i is described by a binary state $x_i(t) \in \{-1, +1\}$ at discrete time $t \in \mathbb{N}$. Choosing the alphabet to be $\{-1, +1\}$ rather than $\{0, 1\}$ as more common in the literature on BNs, will turn out to be advantageous later. However, both choices are equivalent. The state of the network at time t can be described by the vector $\mathbf{x}(t) = \{x_1(t), \dots, x_N(t)\} \in \{-1, +1\}^N$. The network dynamic is defined by

$$x_i(t+1) = f_i(\mathbf{x}(t)), \quad (1)$$

where $f_i : \{-1, +1\}^N \rightarrow \{-1, +1\}$ is the Boolean function associated with node i . At time $t = 0$, an initial state $\mathbf{x}(0) = \mathbf{x}_0$ is chosen. In general not all arguments x_1, \dots, x_N of a function $f_i(\mathbf{x})$ need to be *relevant*. The variable $x_j, j \in \{1, \dots, N\}$ is relevant for f_i if and only if there exists at least one $\mathbf{x} \in \{-1, +1\}^N$ such that changing x_j to $-x_j$ will also change the function’s value. In most of the Boolean network models in biology, the functions depend on a small subset of their arguments only. Furthermore, not every state must have a function associated with it; states can also be external inputs to the network.

To investigate determinative power of nodes and effects of perturbations, we assume that each state is a random variable X_i which follows the distribution $\Pr[X_i = x_i], x_i \in \{-1, +1\}$. We also assume that the random variables X_1, \dots, X_N are independent. This is a natural approach since to study determinative power and tolerance to perturbations, a probabilistic setup is needed. For a given network, the distribution $\Pr[X_i = x_i], x_i \in \{-1, +1\}$ can be estimated by observing the state

$x_i(t)$ sufficiently long. The assumption of independence holds for networks with tree-like structure, but is not feasible for networks with a non-tree like topology with strong local dependencies. In many relevant settings a BN has a tree-like topology, for instance the *E. coli* network which is analyzed in Section 6. For a network with arbitrary topology but few local dependencies, assuming independence will lead to a small modeling error. However, major results concerning the analysis of BNs have been obtained under the assumptions as stated above. For example an important result about the spread of perturbations in random Boolean networks, the annealed approximation [13] has been obtained by assuming that the network size N goes to infinity, which implies there are no local dependencies and the assumptions as stated above apply.

2.2 Notation

We use $[n]$ for the set $\{1, 2, \dots, n\}$, and all sets are subsets of $[n]$. With $\sum_{S \subseteq A}$ we mean the sum over all sets S that are subsets of A . Throughout this paper, we use capital letters for random variables, e.g., X , and lower case letters for their realizations, e.g., x . Boldface letters denote vectors, e.g., \mathbf{X} is a random vector, and \mathbf{x} its realization. For a vector \mathbf{x} and a set $A \subseteq [n]$, \mathbf{x}_A denotes the subvector of \mathbf{x} corresponding to the entries indexed by A .

2.3 Fourier Analysis of Boolean Functions

In the following we give a short introduction to Fourier analysis of Boolean functions $f : \{-1, +1\}^n \rightarrow \{-1, +1\}$. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a binary, product distributed random vector, i.e., the entries of \mathbf{X} are independent random variables $X_i, i \in [n]$ with distribution $\Pr[X_i = x_i], x_i \in \{-1, +1\}$. Then $f(\mathbf{X})$ is a random variable. Throughout this paper, probabilities $\Pr[\cdot]$ and expectations $\mathbb{E}[\cdot]$ are with respect to the distribution of \mathbf{X} . We denote $p_i \triangleq \Pr[X_i = 1]$, the standard deviation of X_i as $\sigma_i \triangleq \sqrt{\text{Var}(X_i)}$, where $\text{Var}(X_i)$ denotes the variance of X_i , and the mean of X_i as $\mu_i \triangleq \mathbb{E}[X_i]$. The inner product of the functions $f, g : \{-1, +1\}^n \rightarrow \{-1, +1\}$ with respect to the distribution of \mathbf{X} is defined as

$$\langle f, g \rangle \triangleq \mathbb{E}[f(\mathbf{X})g(\mathbf{X})] = \sum_{\mathbf{x} \in \{-1, 1\}^n} \Pr[\mathbf{X} = \mathbf{x}] f(\mathbf{x})g(\mathbf{x}) \quad (2)$$

which induces the norm $\|f\| = \sqrt{\langle f, f \rangle}$. An orthonormal basis with respect to the distribution $\Pr[\mathbf{X} = \mathbf{x}]$ is given by the functions

$$\Phi_S(\mathbf{x}) = \prod_{i \in S} \frac{x_i - \mu_i}{\sigma_i}, \quad S \subseteq [n] \setminus \emptyset$$

and

$$\Phi_S(\mathbf{x}) = 1, \quad S = \emptyset.$$

This basis was first proposed by Bahadur [14]. Thus, each function $f : \{-1, +1\}^n \rightarrow \{-1, +1\}$ can be uniquely expressed as

$$f(\mathbf{x}) = \sum_{S \subseteq [n]} \hat{f}(S) \Phi_S(\mathbf{x}), \quad (3)$$

where $\hat{f}(S) \triangleq \langle f, \Phi_S \rangle$ are the Fourier coefficients. Note that (3) is a representation of the function f as a multilinear polynomial, and the Fourier coefficients are the coefficients of that polynomial. As an example consider the AND2 function, i.e., $f_{\text{AND2}}(\mathbf{x}) = 1$ if $x_1 = x_2 = 1$ and $f_{\text{AND2}}(\mathbf{x}) = -1$

for all choices of \mathbf{x} . Suppose the input variables X_1, X_2 are uniformly distributed, i.e., $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$. Then the Fourier coefficients are $\hat{f}(\emptyset) = -1/2$, $\hat{f}(\{1\}) = 1/2$, $\hat{f}(\{2\}) = 1/2$ and $\hat{f}(\{1, 2\}) = 1/2$. Hence (3) becomes

$$f_{AND}(\mathbf{x}) = -\frac{1}{2} + \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_1x_2.$$

As a second example consider the parity function of 2 variables, i.e., the XOR function. PARITY2 is defined as $f_{\text{PARITY2}}(\mathbf{x}) = 1$ if $x_1 = x_2 = 1$ or if $x_1 = x_2 = -1$, and $f_{\text{PARITY2}}(\mathbf{x}) = -1$ for all other choices of \mathbf{x} . It is easily seen, that written as a polynomial, that is $f_{\text{PARITY2}}(\mathbf{x}) = x_1x_2$. Let us conclude this section by listing properties of the Fourier transform which are used frequently throughout this paper.

Decomposition: Let $A \subseteq [n]$ and $S \subset A$. Denote $\bar{S} = A \setminus S$, then

$$\Phi_A(\mathbf{x}) = \Phi_S(\mathbf{x})\Phi_{\bar{S}}(\mathbf{x}).$$

Orthonormality: For $A, B \subseteq [n]$

$$\mathbb{E}[\Phi_A(\mathbf{X})\Phi_B(\mathbf{X})] = \begin{cases} 1, & \text{if } A = B \\ 0, & \text{otherwise.} \end{cases}$$

Parseval's identity: For $f : \{-1, +1\}^n \rightarrow \{-1, +1\}$,

$$\mathbb{E}[f(\mathbf{X})^2] = \|f\|^2 = \sum_{S \subseteq [n]} \hat{f}(S)^2 = 1.$$

3 Influence and Average Sensitivity

In this section we discuss measures of perturbations and their relation to the Fourier spectrum. We start with the *influence* which is a measure of the perturbation of a single input.

Definition 1 ([15]). *Define the influence of variable i on the function f as*

$$I_i(f) = \Pr[f(\mathbf{X}) \neq f(\mathbf{X} \oplus e_i)],$$

where $\mathbf{x} \oplus e_i$ is the vector obtained from \mathbf{x} by flipping its i th entry.

By definition, the influence of variable i is the probability that a perturbation of input i , i.e., flipping input i , changes the function's output. Hence influence captures the effect of a single perturbation of input i . In [15], the authors considered a setting where n processors were agreeing on some common bit as the output of a Boolean function. They used the notion of influence as a measure to capture the probability that a single faulty processor can alter the result of the common bit, if the faulty processor knows the bits announced by all other processors. This illustrates that influence can be viewed as the capability of input i to change the output of function f . In Boolean networks, usually the sum of all influences, i.e., the *average sensitivity* is studied.

Definition 2. The average sensitivity of f to the variables in set A is defined as

$$I_A(f) = \sum_{i \in A} I_i(f).$$

The average sensitivity of f is defined as $as(f) = I_{[n]}(f)$.

$I_A(f)$ captures whether flipping an input, chosen uniformly at random from the set A affects the function's output. Most commonly all inputs are taken into account, i.e., the average sensitivity $as(f)$ is studied, because it measures the effect of the perturbation of a single input. As an example, for the previously introduced AND2 and PARITY2 functions we have $as(f_{PARITY2}) = 2$ and $as(f_{AND2}) = 1$, hence, the PARITY2 function is more sensitive to single perturbations than the AND2 function is. In the following we discuss the Fourier spectrum of functions that are tolerant to perturbations. To this end, the influence and average sensitivity are given in terms of Fourier coefficients as follows.

Proposition 1. For any Boolean function f ,

$$I_i(f) = \frac{1}{\sigma_i^2} \sum_{S \subseteq [n]: i \in S} \hat{f}(S)^2.$$

Proposition 2. For any Boolean function f ,

$$I_A(f) = \sum_{S \subseteq [n]} \hat{f}(S)^2 \sum_{S: i \in S \cap A} \frac{1}{\sigma_i^2} \quad (4)$$

Proposition 1 appears in [16, Lem. 4.1]. Proposition 2 follows directly from Proposition 1 and the definition of $I_A(f)$. From (4) it becomes apparent that the average sensitivity $as(f)$ is large if the Fourier weight, i.e., the squared Fourier coefficient, $\hat{f}(S)^2$, is concentrated on the coefficients of high degree $d = |S|$. Parseval's identity implies, that the terms $\hat{f}(S)^2$ for which the degree $d = |S|$ is small must then be small. Let's see an example which demonstrates this matter: Consider the AND3 function, i.e., $f_{AND3}(x_1, x_2, x_3) = 1$ if and only if $x_1 = x_2 = x_3 = 1$. The average sensitivity of the AND3 function is $as(f_{AND3}) = 0.75$. Hence, f_{AND3} is tolerant to perturbations. In Figure 3, we see that the spectrum of f_{AND3} is concentrated on the coefficients of low degree. In contrast, consider the parity of three variables: $f_{PARITY3}(x_1, x_2, x_3) = x_1 x_2 x_3$, for which $as(f_{PARITY3}) = 3$. Hence, PARITY3 is very sensitive to perturbations. Observe in Figure 3 that the spectrum of the PARITY3 function is concentrated to a maximum on the coefficient of highest degree: $\hat{f}(\{1, 2, 3\}) = 1$.

It can be seen from (4) that for a function f to be tolerant to single perturbations, i.e., to have a small average sensitivity, the Fourier coefficients must be concentrated on coefficients with low degree. When \mathbf{X} is uniform distributed, this is the case if f is strongly biased (i.e. if $f(\mathbf{x}) = a$, for most inputs \mathbf{x} , where $a \in \{-1, 1\}$ is a constant). Then the coefficient $\hat{f}(S)$ of smallest degree $d = |S|$, which is $\hat{f}(\emptyset)$, is large, and the other coefficients must be small according to Parseval's identity. If a function f depends on few variables it follows from (4) that the average sensitivity is small as the degree $d = |S|$ is small for all $\hat{f}(S)$. That implies, a function which depends on few

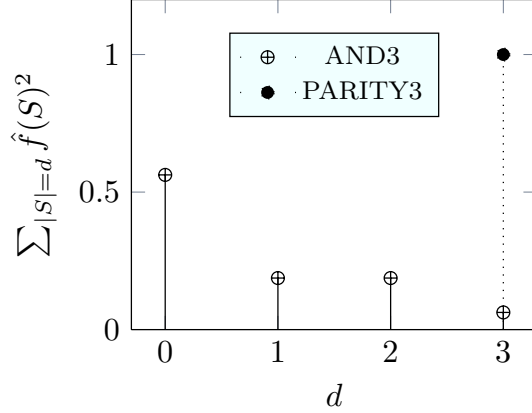


Figure 1: The Fourier spectrum of the AND3 and PARITY3 function.

variables, is tolerant to perturbations, which is in accordance with the observations and results of Kauffman; he found that if a random BN operates in the ordered regime, then the functions in the network depend in average on few variables.

So far, we only discussed single perturbations. However, the discussion also carries over to other noise models, because if a function is tolerant to single perturbations, its *noise sensitivity* is small. The noise sensitivity of a Boolean functions is defined as the probability that a function's output changes if each input is flipped with probability ϵ . For uniform distributed \mathbf{X} , $\epsilon as(f)$ is an upper bound for the noise sensitivity, and for small values of ϵ , $\epsilon as(f)$ actually approximates the noise sensitivity well. For a different distribution of \mathbf{X} and a slightly different noise model Schober [17] found that $\epsilon as(f)$ still upper bounds the noise sensitivity. This result was generalized in [18].

4 Mutual Information and Uncertainty

In this section, we study the mutual information $MI(f(\mathbf{X}); \mathbf{X}_A)$ between $f(\mathbf{X})$ and a subset of variables \mathbf{X}_A , where \mathbf{X}_A consists of the entries of \mathbf{X} corresponding to the indices in the set $A \subseteq [n]$. We argue that $MI(f(\mathbf{X}); \mathbf{X}_A)$ is a measure of the determinative power of \mathbf{X}_A over $f(\mathbf{X})$. If $MI(f(\mathbf{X}); \mathbf{X}_A) = 0$, then \mathbf{X}_A and $f(\mathbf{X})$ are statistically independent even when \mathbf{X}_A are relevant variables. We will characterize functions for which this is the case in terms of Fourier coefficients. This provides insight for which functions sets of inputs and the function's output are statistically dependent.

Before giving a formal definition of mutual information, let us start with the following example. Consider the PARITY2 function where the inputs X_1, X_2 are uniformly distributed. Intuitively, if X_1 has determinative power, knowledge about X_1 should provide us with information about the random variable $f_{\text{PARITY2}}(\mathbf{X})$. Suppose we know the value of X_1 , say $X_1 = 1$. Since $f_{\text{PARITY2}}(\mathbf{X}) = X_1 X_2$, we have with $\Pr[X_2 = 1] = 1/2$ that $\Pr[f_{\text{PARITY2}}(\mathbf{X}) = 1] = \Pr[f_{\text{PARITY2}} = 1 | X_1 = 1]$. Hence, knowing the value of X_1 does not help to predict the value of $f_{\text{PARITY2}}(\mathbf{X})$. Therefore X_1 has no determinative power over $f_{\text{PARITY2}}(\mathbf{X})$. We indeed have $MI(f_{\text{PARITY2}}(\mathbf{X}); X_1) = 0$.

Let us now define mutual information. Mutual information is the reduction of uncertainty of a random variable Y due to the knowledge of X ; therefore we need to define a measure of uncertainty first which is the entropy of a random variable. As a reference for the following definitions see [19].

Definition 3. The entropy $H(X)$ of a discrete random variable X with alphabet \mathcal{X} is defined as

$$H(X) \triangleq - \sum_{x \in \mathcal{X}} \Pr[X = x] \log_2 \Pr[X = x].$$

Definition 4. The conditional entropy $H(Y|X)$ of a pair of discrete and jointly distributed random variables (Y, X) is defined as

$$H(Y|X) \triangleq \sum_{x \in \mathcal{X}} \Pr[X = x] H(Y|X = x).$$

Definition 5. The mutual information $\text{MI}(Y; X)$ is the reduction of uncertainty of the random variable Y due to knowledge of X ,

$$\text{MI}(Y; X) \triangleq H(Y) - H(Y|X).$$

For a binary random variable X with $p \triangleq \Pr[X = 1]$, we have $H(X) = h(p)$, where $h(p)$ is the binary entropy function, defined as

$$h(p) \triangleq -p \log_2 p - (1 - p) \log_2 (1 - p). \quad (5)$$

Mutual information is a measure of determinative power because of the following reasons. Consider a single variable X_i of the argument \mathbf{X} : If knowledge of X_i reduces the uncertainty of $f(\mathbf{X})$, then X_i determines the state of $f(\mathbf{X})$ to some extent, because then knowledge about the state of X_i helps in predicting $f(\mathbf{X})$. Another property of mutual information, which we expect from a measure of determinative power is that not all variables can have large determinative power because of the inequality

$$\sum_{i=1}^n \text{MI}(f(\mathbf{X}); X_i) \leq \text{MI}(f(\mathbf{X}); \mathbf{X}) \leq 1, \quad (6)$$

which follows from the chain rule of mutual information (as a reference see [19]) and independence of the variables $X_i, i \in [n]$. Hence, if $\text{MI}(f(\mathbf{X}); X_i)$ is large, i.e., close to 1, we can be sure that X_i has some determinative power over $f(\mathbf{X})$, since (6) implies that $\text{MI}(f(\mathbf{X}); X_j)$ for $j \neq i$ must be small then. That is one reason why influence is not a measure of determinative power: Each value can have large influence, consider e.g. the parity function, where each input has influence 1. If variable i has large influence, this just implies that input i has power to change the output, but not to determine it.

In the following we represent the mutual information in terms of the Fourier coefficients and then relate the mutual information to the influence. We first consider single variables, i.e., $\text{MI}(f(\mathbf{X}); X_i)$, and then the general case of a set of variables, i.e., $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$.

4.1 Single Variables

$\text{MI}(f(\mathbf{X}); X_i)$ has previously been studied as *information gain* as a measure of “goodness” for split variables in greedy tree learners [20], and also appears under the name *informativeness* as a measure

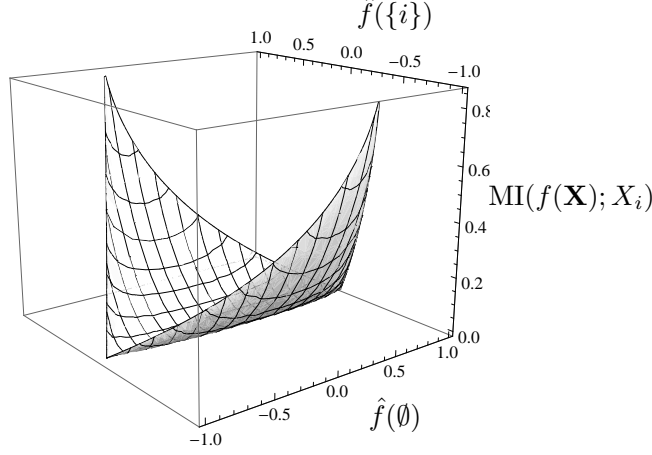


Figure 2: $\text{MI}(f(\mathbf{X}); X_i)$ as a function of $\hat{f}(\{i\})$ and $\hat{f}(\emptyset)$ for $p_i = 0.3$.

of voting power [21]. First, we discuss $\text{MI}(f(\mathbf{X}); X_i)$ and its relation to the influence. We start by expressing the mutual information in terms of Fourier coefficients as

$$\begin{aligned} \text{MI}(f(\mathbf{X}); X_i) = & h\left(\frac{1}{2}(1 + \hat{f}(\emptyset))\right) \\ & - p_i h\left(\frac{1}{2}\left(1 + \hat{f}(\emptyset) + \hat{f}(\{i\})\frac{1 - \mu_i}{\sigma_i}\right)\right) \\ & - (1 - p_i) h\left(\frac{1}{2}\left(1 + \hat{f}(\emptyset) + \hat{f}(\{i\})\frac{-1 - \mu_i}{\sigma_i}\right)\right), \end{aligned} \quad (7)$$

which follows from Theorem 5, given in Section 4.3. The mutual information $\text{MI}(f(\mathbf{X}); X_i)$ just depends on $\hat{f}(\{i\})$, $\hat{f}(\emptyset)$ and p_i . In contrast, the influence $I_i(f)$ is a function of the coefficients $\{\hat{f}(S) : S \in [n], i \in S\}$. In Figure 2 we depict $\text{MI}(f(\mathbf{X}); X_i)$ for $p_i = 0.3$ as a function of $\hat{f}(\{i\})$ and $\hat{f}(\emptyset)$. It can be seen that $\text{MI}(f(\mathbf{X}); X_i) = 0$, i.e., $f(\mathbf{X})$ and X_i are statistically independent, if and only if $\hat{f}(\{i\}) = 0$. That can be formalized as follows. $\text{MI}(f(\mathbf{X}); X_i)$ is convex in $\hat{f}(\{i\})$. This can be proven by taking the second derivative of (7), and observing that it is larger than zero for all pairs of values $(\hat{f}(\emptyset), \hat{f}(\{i\}))$ for which $\text{MI}(f(\mathbf{X}); X_i)$ is defined. Next, from (7) we see that $\text{MI}(f(\mathbf{X}); X_i) = 0$ if $\hat{f}(\{i\}) = 0$; hence it follows that $\text{MI}(f(\mathbf{X}); X_i) = 0$ if and only if $\hat{f}(\{i\}) = 0$, which proves the following corollary.

Corollary 1. *Let f be a Boolean function and \mathbf{X} be product distributed. X_i and $f(\mathbf{X})$ are statistically independent if and only if $\hat{f}(\{i\}) = 0$.*

Corollary 1 also follows from a more general result, Theorem 3, which is presented later.

We already verified Corollary 1 in the introductory example: For the PARITY2 function we recognized

$\text{MI}(f_{\text{PARITY2}}(\mathbf{X}); X_1) = 0$ and $\hat{f}(\{1\}) = 0$. From Figure 2 it can be seen that the larger $|\hat{f}(\{i\})|$, the larger $\text{MI}(f(\mathbf{X}); X_i)$ becomes. Formally, it follows from the convexity and Corollary 1, that $\text{MI}(f(\mathbf{X}); X_i)$ is increasing in $|\hat{f}(\{i\})|$. Hence X_i has large determinative power, i.e., $\text{MI}(f(\mathbf{X}); X_i)$ is large, if and only if $|\hat{f}(\{i\})|$ is large (i.e., close to one). $|\hat{f}(\{i\})|$ is trivially maximized for the

dictatorship function, i.e., for $f(\mathbf{x}) = x_i$, or its negation, i.e., $f(\mathbf{x}) = -x_i$. The output $f(\mathbf{X})$ of the dictatorship function is fully determined by x_i .

Let us now relate the mutual information to the influence:

Corollary 2. *For any Boolean function f , for any product distributed \mathbf{X} ,*

$$I_i(f) \geq \frac{1}{\sigma_i^2} (\text{MI}(f(\mathbf{X}); X_i) - \Psi(\text{Var}(f(\mathbf{X})))) ,$$

with $\Psi(x) = (x)^{1/\ln(4)} - x$.

Proof. Follows from Theorem 2 given in Section 4.2. □

The term $\Psi(\text{Var}(f(\mathbf{X})))$ should be understood as an error term which satisfies $0 \leq \Psi(\text{Var}(f(\mathbf{X}))) < 0.12$ and which is close to zero for settings of interest, as the following argument explains. Corollary 2 is not of interest when $\text{Var}(f(\mathbf{X}))$ is small, since then $f(\mathbf{X})$ is close to a constant function (i.e., close to $f(\mathbf{X}) = 1$ or $f(\mathbf{X}) = -1$), and $I_i(f)$ and $\text{MI}(f(\mathbf{X}); X_i)$ must both be small, i.e., close to zero. Hence Corollary 2 is of interest when $\text{Var}(f(\mathbf{X}))$ is large, i.e., close to 1, and then the term $\Psi(\text{Var}(f(\mathbf{X})))$ is small, e.g., for $\text{Var}(f(\mathbf{X})) > 0.8$, $\Psi(\text{Var}(f(\mathbf{X}))) < 0.05$. Corollary 2 gives a lower bound on the influence of a variable by the mutual information of that variable. Hence, if $\text{MI}(f(\mathbf{X}); X_i)$ is large, then $I_i(f)$ is also large. That proves the intuitive idea that if an input determines $f(\mathbf{X})$ to some extent, this input also has to be sensitive to errors. Conversely, as mentioned previously, an input i can have large influence and still $\text{MI}(f(\mathbf{X}); X_i) = 0$: For the PARITY2 function we have $I_i(f) = 1$ and $\text{MI}(f(\mathbf{X}); X_i) = 0$.

Interestingly, the influence also has an information theoretic interpretation. The following theorem generalizes Theorem 1 in [21].

Theorem 1. *For any Boolean function f , for any product distributed \mathbf{X} ,*

$$I_i(f) = \frac{H(f(\mathbf{X}) | \mathbf{X}_{[n] \setminus \{i\}})}{H(X_i)}.$$

Proof. See Appendix B. For uniform distributed \mathbf{X} , a proof appears in [21]. □

Proposition 1 shows that the influence of a variable is a measure for the uncertainty of the function's output that remains if all variables except variable i are set.

4.2 A Set of Variables

In this part we discuss $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$ where A is an arbitrary subset of $[n]$.

Let us first consider the case where $A = [n]$. Then $\mathbf{X}_A = \mathbf{X}$ and $\text{MI}(f(\mathbf{X}); \mathbf{X}_A) = H(f(\mathbf{X}))$. It follows that $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$ is maximized for $\Pr[f(\mathbf{X}) = 1] = 1/2$, i.e., if the variance of $f(\mathbf{X})$ is 1. This is the case if $\hat{f}(\emptyset) = 0$. In general, the closer to zero $\hat{f}(\emptyset)$ is, the larger the mutual information between a function's output and all its inputs.

From now on, let A be an arbitrary subset of $[n]$. In the following we relate the average sensitivity of the inputs indexed by the set A to the mutual information.

Theorem 2. For any Boolean function f , for any product distributed \mathbf{X} ,

$$I_A(f) \geq \min_{i \in A} \left(\frac{1}{\sigma_i^2} \right) (\text{MI}(f(\mathbf{X}); \mathbf{X}_A) - \Psi(\text{Var}(f(\mathbf{X})))) \quad (8)$$

with

$$\Psi(x) \triangleq (x)^{1/\ln(4)} - x. \quad (9)$$

Proof. See Appendix C. □

As explained previously, the term $\Psi(\text{Var}(f(\mathbf{X})))$ should be understood as an error term. Theorem 2 shows that a large value of $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$ implies that f must be sensitive to perturbations of the variables \mathbf{X}_A . Moreover, if $I_A(f)$ is small, i.e., if f is tolerant to perturbations of the variables \mathbf{X}_A , then $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$ must be small, i.e., the variables \mathbf{X}_A do not have large determinative power. For the case that $A = [n]$, Theorem 2 states that the average sensitivity is lower-bounded by $\text{MI}(f(\mathbf{X}); \mathbf{X})$ minus some small term.

The following theorem characterizes statistical independence of $f(\mathbf{X})$ and a set of its arguments \mathbf{X}_A in terms of the Fourier coefficients. This result generalizes a theorem derived by Xiao and Massey [10] from uniform to product distributed \mathbf{X} .

Theorem 3. Let $A \subseteq [n]$ be fixed, f be a Boolean function and \mathbf{X} be product distributed. Then $f(\mathbf{X})$ and the inputs $\mathbf{X}_A = \{X_i : i \in A\}$ are statistically independent if and only if

$$\hat{f}(S) = 0 \text{ for all } S \subseteq A \setminus \emptyset.$$

Proof. See Appendix D. For uniform distributed \mathbf{X} , i.e., $\Pr[X_i = 1] = 1/2 \forall i \in [n]$, Theorem 3 has been derived by Xiao and Massey [10]. The proof provided here follows from an application of Lemma 1 in Appendix A, whereas the proof for uniform distributed \mathbf{X} given in [10] relies on the Xiao-Massey lemma. □

Theorem 3 shows that a function and small sets of its inputs are statistically independent, if the spectrum is concentrated on the coefficients of high degree $d = |S|$. The most prominent example is the parity function of n variables, i.e., $f_{\text{PARITYN}}(\mathbf{x}) = x_1 x_2 \dots x_n$: For uniformly distributed \mathbf{X} , each subset of $n - 1$ or fewer arguments and $f_{\text{PARITYN}}(\mathbf{X})$ are statistically independent. If a function is concentrated on the coefficients of low degree $d = |S|$, which is the case for functions that are tolerant to perturbations, then small sets of inputs and the function's output are statistically dependent.

Theorem 3 is also of interest to analyze algorithms which detect functional dependencies in a BN based on estimating the mutual information from observations of the network's states, such as the algorithm presented in [11]. For those, Theorem 3 allows to predict for which classes of functions such an algorithm can succeed and for which it will fail. Theorem 3 also shows that in a BN model of a genetic regulatory network, a functional dependency between a gene and a regulator cannot be detected based on statistical dependence of a regulator X_i and a gene's state $f_j(\mathbf{X})$, unless the model restricts the regulatory functions to those for which $\hat{f}(\{i\}) > 0$ holds for each function's input.

4.3 Fourier Coefficients and Uncertainty

Let us finally relate the conditional entropy $H(f(\mathbf{X})|\mathbf{X}_A)$ to a concentration of the Fourier weight on the coefficients $\{S: S \subseteq A\}$ where $A \subseteq [n]$.

Theorem 4. *Let f be a Boolean function, let \mathbf{X} be product distributed and let $\mathbf{X}_A = \{X_i: i \in A\}$ be a fixed set of arguments, where $A \subseteq [n]$. Then*

$$\left(1 - \sum_{S \subseteq A} \hat{f}(S)^2\right)^{\frac{1}{\ln(4)}} \geq H(f(\mathbf{X})|\mathbf{X}_A) \geq 1 - \sum_{S \subseteq A} \hat{f}(S)^2.$$

Proof. See Appendix E. □

Theorem 4 shows that $H(f(\mathbf{X})|\mathbf{X}_A)$ can be approximated with $1 - \sum_{S \subseteq A} \hat{f}(S)^2$. It further shows that $H(f(\mathbf{X})|\mathbf{X}_A)$ is small, if the Fourier weight is concentrated on the variables in the set A , i.e., if $\sum_{S \subseteq A} \hat{f}(S)^2$ is close to one. The conditional entropy $H(f(\mathbf{X})|\mathbf{X}_A)$ can also be directly expressed in terms of Fourier coefficients:

Theorem 5. *Let f be a Boolean function, let \mathbf{X} be product distributed and let $\mathbf{X}_A = \{X_i: i \in A\}$ be a fixed set of arguments, where $A \subseteq [n]$. Then*

$$H(f(\mathbf{X})|\mathbf{X}_A) = \mathbb{E} \left[h \left(\frac{1}{2} \left(1 + \sum_{S \subseteq A} \hat{f}(S) \Phi_S(\mathbf{X}_A) \right) \right) \right]$$

where $h(\cdot)$ is the binary entropy function as defined in (5).

Proof. See Appendix F. For the special case of uniform distributed \mathbf{X} , a proof appears in [22], in the the context of designing S-Boxes. □

Theorem 5 shows that the conditional entropy $H(f(\mathbf{X})|\mathbf{X}_A)$ is a function of the coefficients $\{\hat{f}(S): S \subseteq A\}$ only. In contrast the average sensitivity depends on all coefficients $\{\hat{f}(S): |S \cap A| > 0\}$.

5 Unate Functions

In this section we discuss unate, i.e., locally monotone functions.

Definition 6. *A Boolean function f is said to be unate in x_i if for each $\mathbf{x} = (x_1, \dots, x_n) \in \{-1, +1\}^n$ and for some fixed $a_i \in \{-1, +1\}$, $f(x_1, \dots, x_i = -a_i, \dots, x_n) \leq f(x_1, \dots, x_i = a_i, \dots, x_n)$ holds. f is said to be unate, if f is unate in each variable x_i , $i \in [n]$.*

Each linear threshold function is unate and also most, if not all, regulatory interactions in a biological network are considered to be unate. That can be deduced from [12, 23], and the basic argument is the following: If an element acts either as a repressor or an activator for some gene, but never as both, then the function determining the gene's state is unate by definition. And this is an reasonable assumption for regulatory interactions [12, 23]. For unate functions, the following interesting property holds.

Proposition 3. Let $f : \{-1, +1\}^n \rightarrow \{-1, +1\}$ be unate. Then

$$\hat{f}(\{i\}) = a_i \sigma_i I_i(f), \forall i \in [n] \quad (10)$$

where $a_i \in \{-1, +1\}$ is the parameter as given in Definition 6.

Proof. Goes along the same lines as the proof for monotone functions in [16, Lem. 4.5]. \square

Note that conversely, if (10) holds for each $x_i, i \in [n]$, f must not be unate. Inserting (10) into (7) yields

$$\begin{aligned} \text{MI}(f(\mathbf{X}); X_i) &= h\left(\frac{1}{2}(1 + \hat{f}(\emptyset))\right) \\ &\quad - \mathbb{E}\left[h\left(\frac{1}{2}\left(1 + \hat{f}(\emptyset) + a_i \sigma_i I_i(f) \frac{X_i - \mu_i}{\sigma_i}\right)\right)\right] \end{aligned} \quad (11)$$

where the expectation in (11) is over X_i . Based on (11), the discussion from Section 4.1 applies by using $\hat{f}(\{i\})$ and $a_i \sigma_i I_i(f)$ synonymously. Hence, for unate functions, the mutual information $\text{MI}(f; X_i)$ is increasing in the influence $I_i(f)$. Moreover if f is unate, and x_i is a relevant variable, i.e., a variable on which the functions actually depends on, then $|\hat{f}(\{i\})| > 0$. From this fact and the same arguments as given in Section 4.1 follows:

Theorem 6. Let $f : \{-1, +1\}^n \rightarrow \{-1, +1\}$ be unate. If and only if x_i is a relevant variable, then $\text{MI}(f(\mathbf{X}); X_i) \neq 0$.

In a Boolean model of a biological regulatory network, this implies that if the functions in the network are unate, then a regulator and the target gene must be statistically dependent.

6 *E. coli* Regulatory Network

In [6], the authors presented a complex computational model of the *E. coli* transcriptional regulatory network that controls central parts of the *E. coli* metabolism. The network consists of 798 nodes and 1160 edges. 636 of the nodes represent genes and of the remaining 162 ones, most (103) are external metabolites. The rest are stimuli and other state variables such as internal metabolites. The network has a layered feed-forward structure, i.e., no feedback-loops exists. The 133 elements in the first layer can be viewed as the inputs of the system and the elements in the following 7 layers are interacting genes representing the internal state of the system. Our investigations showed that all functions are unate. This is a special property of the network, since if functions are chosen uniformly at random, it is unlikely to sample a unate function, especially if the degree n is large. Hence, the properties which we derived in Section 5 apply.

6.1 Determinative Nodes in the *E. coli* Network

In this section, we identify the input-nodes that have large determinative power (we will define what that means shortly), and show that a small number thereof reduces the uncertainty about the network's state significantly. More specifically, we show that in average the entropy of the controlled states, conditioned on a small set of determinative input nodes, is small. That implies

that the conditional probability of the controlled nodes to be either one or minus one, is close to one in average. We denote by $\mathbf{X} = \{X_1, \dots, X_n\}$, $n = 162$ the set of inputs of the feed-forward network and assume they are independent and uniformly distributed. The remaining variables are denoted by $\mathbf{Y} = \{Y_1, \dots, Y_m\}$, $m = 636$, and are given as a function of the input nodes, and also of other networks states, i.e., $Y_i = f'_i(\mathbf{X}, \mathbf{Y})$. For the analysis, the distributions of the random variables Y_1, \dots, Y_m need to be computed, since some of these variables are arguments to other functions. This can be circumvented by defining a collapsed network, i.e., a network where each state in the network is given as a function of the input nodes, i.e., $Y_i = f_i(\mathbf{X})$. The collapsed network is obtained by inserting consecutively functions into each other, until each function just depends on the states of the nodes in the input layer, i.e., on \mathbf{X} . After having obtained each state Y_i as a function of the input \mathbf{X} , the determinative nodes can be identified.

As argued in Section 4, $\text{MI}(f_i(\mathbf{X}); X_j)$ is a measure of the determinative power of X_j over $Y_i = f_i(\mathbf{X})$. This motivates to define the determinative power of input X_j over the states in the network as:

$$D(j) \triangleq \sum_{i=1}^m \text{MI}(f_i(\mathbf{X}); X_j).$$

Note that a small value of $D(j)$ just implies that the state X_j alone does not have large determinative power over states in the network. However, $\sum_{i=1}^m \text{MI}(f_i(\mathbf{X}); X_j, X_k)$ can be large for some $j, k \in [n]$, even though $D(j)$ and $D(k)$ are both small. We computed $D(j)$ for each input variable and found that $D(j)$ is large just for some inputs, such as the variables *o2_xt* (36.9 bit), *leu-l_xt* (20.9 bit) and *glc-d_xt* (19.3 bit), (here we adopted the names from the original dataset), but is small for most other variables. It is natural to ask, whether this can be explained from nodes with large values of $D(j)$ having many outgoing edges, while most other nodes do not, i.e., the out-degree of the input nodes is power law distributed. Indeed, in the *E. coli* network, the out-degree and $D(j)$ are correlated to some extent. However, in theory, having an large out-degree does not necessarily result in large values of $D(j)$. That is also what we observe in practice, e.g., the state variable *glc-d_xt* has 99 outgoing edges, but $D(\text{glc-d_xt}) = 19.3 \text{ bit}$, whereas variable *o2_xt* has out degree 72, but $D(\text{o2_xt}) = 36.9 \text{ bit}$. For the following, denote by τ a permutation on $[n]$ such that $D(X_{\tau(1)}) \geq D(X_{\tau(2)}) \geq \dots \geq D(X_{\tau(n)})$, i.e., τ orders the inputs nodes in descending order in their determinative power.

After having identified the most determinative nodes, we want to see whether knowledge about a small set of those reduces the entropy of the networks states significantly, i.e., we are interested in $H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)})$ as a function of l . The quantity $H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)})$ has an interesting interpretation which arises as a consequence of the so called *asymptotic equipartition property*, see [19]: Consider a sequence $\mathbf{y}_1, \dots, \mathbf{y}_k$ of k samples of the random variable \mathbf{Y} . For $\epsilon > 0$ and k sufficiently large, there exists a set $A_\epsilon^{(k)}$ of typical sequences $\mathbf{y}_1, \dots, \mathbf{y}_k$ such that

$$|A_\epsilon^{(k)}| \leq 2^{k(H(\mathbf{Y})+\epsilon)}$$

and

$$\Pr \left[\mathbf{Y} \in A_\epsilon^{(q)} \right] > 1 - \epsilon$$

where $|A_\epsilon^{(k)}|$ denotes the cardinality of the set $A_\epsilon^{(k)}$. Namely, the sequences obtained by drawing from \mathbf{Y} are likely to fall in a set of size determined by the uncertainty of \mathbf{Y} . In this sense

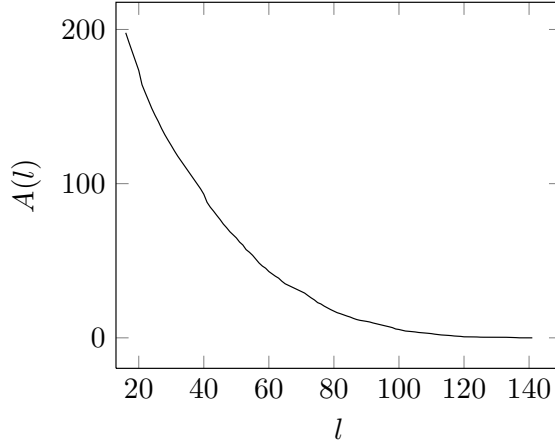


Figure 3: The upper bound $A(l)$ on $H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)})$ as a function of l .

$H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)})$ can be interpreted as a measure of the size of a subset of the overall state space where the system is likely to be found, given knowledge about the states $X_{\tau(1)}, \dots, X_{\tau(l)}$.

For a large network, $H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)})$ is hard to compute directly, therefore we upper bound this quantity. The random variable \mathbf{Y} is a function of \mathbf{X} , hence its entropy cannot be larger than the entropy of \mathbf{X} and we obtain the simple upper bound

$$H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)}) \leq H(\mathbf{Y}) \leq H(\mathbf{X}) = 162 \text{bit}.$$

We consider the upper bound

$$H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)}) \leq A(l) \triangleq \sum_{i=1}^m H(Y_i|X_{\tau(1)}, \dots, X_{\tau(l)}).$$

which follows from the chain rule for entropy [19]. The upper bound $A(l)$ on $H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)})$ is depicted in Figure 3 as a function of l . We can see from Figure 3 that knowledge of the states of the most determinative nodes reduces the uncertainty about the network's states significantly. Actually, this bound is not very tight, hence we can even expect $H(\mathbf{Y}|X_{\tau(1)}, \dots, X_{\tau(l)})$ to lie significantly below this upper bound. This experiment demonstrates that the determinative power is unequally distributed among the input nodes in the *E. coli* network. Another interpretation of the bound $A(l)$ is the following. When $A(l)$ is small, on average $H(Y_i|X_{\tau(1)}, \dots, X_{\tau(l)})$ must be small and hence $\Pr[Y_i = 1|X_{\tau(1)}, \dots, X_{\tau(l)}]$ is close to one or to zero in average.

6.2 Tolerance to Perturbations

Finally, we discuss the average sensitivity of individual functions in the *E. coli* network. In Section 3, we found that the average sensitivity is small if the Fourier spectrum is concentrated on the coefficients of low degrees, and that there are basically two types of functions for which this is the case: Functions for which the bias is high, i.e., the probability $\Pr[f(\mathbf{X}) = 1]$ is close to one or zero, and functions that depend just on a few variables. Figure 4 shows pairs of values $(as(f), \Pr[f(\mathbf{X}) = 1])$ for each function in the *E. coli* network, assuming that X_1, \dots, X_n are independent and $\Pr[X_i = 1] = 1/2$. Observe that the functions with high in-degree K (i.e., number of

relevant input variables) could have a large average sensitivity, but $as(f)$ is small, because these functions are highly biased. Hence we can see from Figure 4 that the average sensitivity of the functions is small—indeed close to the lower bound on the average sensitivity—because a function either depends on few coordinates or is highly biased. For other distributions on the inputs, i.e., other values of $p = \Pr[X_i = 1], \forall i \in [n]$, Figure 4 looks similar.

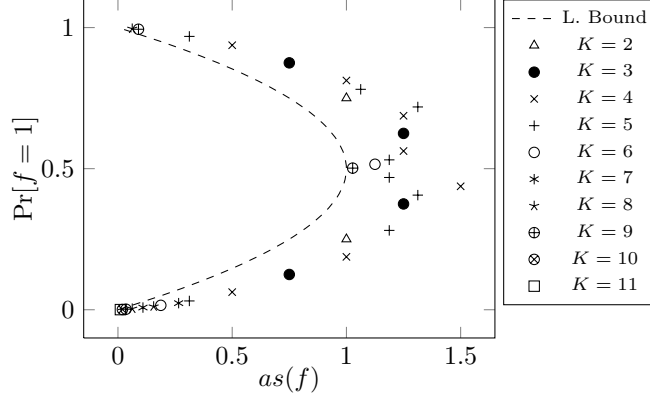


Figure 4: Pairs of values $(as(f), \Pr[f(\mathbf{X}) = 1])$ of each function in the *E. coli* network, for different in-degrees K and uniform distributed \mathbf{X} . Moreover the lower bound on the average sensitivity $as(f)$ is plotted (known as Poincaré's inequality).

7 Conclusion

In a Boolean network, properties such as tolerance to perturbations and statistical dependencies between nodes are properties of single functions. Hence, we concentrated on the analysis of single functions and considered a function $f(\mathbf{X})$ as a random variable, where \mathbf{X} is a product distributed argument. For this setting, we used Fourier analysis of Boolean functions to investigate the mutual information $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$ between a function $f(\mathbf{X})$ and a set of variables \mathbf{X}_A . We argued that $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$ is a measure of the determinative power of \mathbf{X}_A . For the mutual information between a single variable X_i and $f(\mathbf{X})$, we found that $\text{MI}(f(\mathbf{X}); X_i)$ just depends on $(\hat{f}(\{i\}), \hat{f}(\emptyset))$, and is strictly increasing in $\hat{f}(\{i\})$. Furthermore if $\text{MI}(f(\mathbf{X}); X_i)$ is large, the influence $I_i(f)$ must be large. Moreover, we proved inequalities that relate $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$ to measures of perturbations, and gave the necessary and sufficient conditions for statistical independence of $f(\mathbf{X})$ and \mathbf{X}_A . $\text{MI}(f(\mathbf{X}); \mathbf{X}_A)$ just depends on the coefficients $\{\hat{f}(S) : S \subseteq A\}$, whereas the average sensitivity of the variables in A , i.e., a measure of perturbation of the variables in A , depend on $\{\hat{f}(S) : |S \cap A| > 0\}$, which is a fundamental difference. However, for the class of unate functions, which are especially interesting for biological networks, we found a direct relation between MI and influence. For unate functions, $\text{MI}(f(\mathbf{X}); X_i) > 0$ for each relevant input. Therefore, in a biological network such as the *E. coli* regulatory network, where each function is unate, a gene and its regulator must be statistically dependent. As an application of our methods, we analyzed the large-scale regulatory network of *E. coli*. We identified the most *determinative* nodes, and demonstrated that knowledge about the states of a small subset of them reduces the uncertainty about the states of all the other nodes significantly. Furthermore we showed that the *E. coli* network is tolerant to perturbations of its

inputs, and that this can be explained by the Fourier spectrum of the functions in the network being concentrated on the coefficients of low degree.

Acknowledgments

We would like to thank Sara Al-Sayed and Dejan Lazich for helpful discussions and careful reading of the manuscript.

Appendices

A Lemma 1

For the proof of Theorem 3 and 5, we will need the following lemma.

Lemma 1. *Let f be a Boolean function, let \mathbf{X} be product distributed and let $A \subseteq [n]$ and some fixed $\mathbf{x}_A \in \{-1, +1\}^{|A|}$ be given. Then*

$$\mathbb{E}[f(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A] = \sum_{S \subseteq A} \hat{f}(S) \Phi_S(\mathbf{x}_A). \quad (12)$$

Proof. Inserting the Fourier expansion of $f(\mathbf{X})$ given by (3) in the left-hand side of (12) and utilizing the linearity of conditional expectation yields

$$\mathbb{E}[f(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A] = \sum_{S \subseteq [n]} \hat{f}(S) \mathbb{E}[\Phi_S(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A].$$

For $S \subseteq A$,

$$\mathbb{E}[\Phi_S(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A] = \Phi_S(\mathbf{x}_A).$$

Conversely for $S \not\subseteq A$

$$\mathbb{E}[\Phi_S(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A] = 0.$$

To see this assume without loss of generality that $S = A \cup \{j\}$ and $j \notin A$. Using the decomposition property of the basis function as given in Section 2.3,

$$\begin{aligned} \mathbb{E}[\Phi_S(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A] &= \mathbb{E} \left[\prod_{i \in S} \Phi_{\{i\}}(\mathbf{X}) | \mathbf{X}_A = \mathbf{x}_A \right] \\ &= \prod_{i \in S} \mathbb{E}[\Phi_{\{i\}}(\mathbf{X}) | \mathbf{X}_A = \mathbf{x}_A] \end{aligned}$$

which is equal to zero as

$$\mathbb{E}[\Phi_{\{j\}}(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A] = \mathbb{E}[\Phi_{\{j\}}(\mathbf{X})] = 0.$$

□

B Proof of Theorem 1

For notational convenience, let $A = [n] \setminus \{i\}$. By definition of the conditional entropy,

$$\begin{aligned} H(f(\mathbf{X})|\mathbf{X}_A) &= \sum_{\mathbf{x}_A \in \{-1,1\}^{|A|}} \Pr[\mathbf{X}_A = \mathbf{x}_A] H(f(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A) \\ &= \sum_{\mathbf{x}_A \in \{-1,1\}^{|A|}} \Pr[\mathbf{X}_A = \mathbf{x}_A] h(\Pr[f(\mathbf{X}) = 1|\mathbf{X}_A = \mathbf{x}_A]) \end{aligned} \quad (13)$$

where $h(\cdot)$ is the binary entropy function as defined in (5). Observe that

$$h(\Pr[f(\mathbf{X}) = 1|\mathbf{X}_A = \mathbf{x}_A]) = h(\Pr[X_i = 1])$$

if

$$\begin{aligned} &f(X_1 = x_1, \dots, X_i = 1, \dots, X_n = x_n) \\ &\neq f(X_1 = x_1, \dots, X_i = -1, \dots, X_n = x_n) \end{aligned}$$

and

$$h(\Pr[f(\mathbf{X}) = 1|\mathbf{X}_A = \mathbf{x}_A]) = 0$$

otherwise. Hence (13) becomes

$$H(f(\mathbf{X})|\mathbf{X}_A) = \sum_{\mathbf{x}_A \in \{-1,1\}^{|A|}} \Pr[\mathbf{X}_A = \mathbf{x}_A] h(\Pr[X_i = 1]) \mathbf{1}_{\{f(\mathbf{x}) \neq f(\mathbf{x} \oplus e_i)\}}$$

where $\mathbf{x} \oplus e_i$ is the vector obtained from \mathbf{x} by flipping its i th entry, and Theorem 1 follows by using the definition of the influence.

C Proof of Theorem 2

According to Proposition 2,

$$\begin{aligned} I_A(f) &= \sum_{S \subseteq [n]} \hat{f}(S)^2 \sum_{i \in S \cap A} \frac{1}{\sigma_i^2} \\ &\geq \sum_{S \subseteq [n] \setminus \emptyset} \hat{f}(S)^2 |S \cap A| \min_{i \in A} \left(\frac{1}{\sigma_i^2} \right) \\ &\geq \min_{i \in A} \left(\frac{1}{\sigma_i^2} \right) \sum_{S \subseteq A \setminus \emptyset} \hat{f}(S)^2. \end{aligned} \quad (14)$$

Next, we rewrite the lower bound on $H(f(\mathbf{X})|\mathbf{X}_A)$ given by Theorem 4 as

$$\sum_{S \subseteq A \setminus \emptyset} \hat{f}(S)^2 \geq 1 - \hat{f}(\emptyset)^2 - H(f(\mathbf{X})|\mathbf{X}_A). \quad (15)$$

By adding $H(f(\mathbf{X})) - H(f(\mathbf{X}))$ on the right hand side of (15), and using the definition of MI, (15) becomes

$$\sum_{S \subseteq A \setminus \emptyset} \hat{f}(S)^2 \geq \text{MI}(f(\mathbf{X}); \mathbf{X}_A) - H(f(\mathbf{X})) + 1 - \hat{f}(\emptyset)^2. \quad (16)$$

With $\text{Var}(f(\mathbf{X})) = 1 - \hat{f}(\emptyset)^2$ and by using the inequality $H(f(\mathbf{X})) \leq (\text{Var}(f(\mathbf{X})))^{1/\ln(4)}$, given in [24, Thm 1.2], (16) becomes

$$\sum_{S \subseteq A \setminus \emptyset} \hat{f}(S)^2 \geq \text{MI}(f(\mathbf{X}); \mathbf{X}_A) - \Psi(\text{Var}(f(\mathbf{X}))), \quad (17)$$

with $\Psi(\cdot)$ as defined in (9). Finally, Theorem 2 follows by combining (14) and (17).

D Proof of Theorem 3

By definition, $f(\mathbf{X})$ and \mathbf{X}_A are statistically independent if and only if for all $\mathbf{x}_A \in \{-1, +1\}^{|A|}$

$$\Pr[f(\mathbf{X}) = 1 | \mathbf{X}_A = \mathbf{x}_A] = \Pr[f(\mathbf{X}) = 1]. \quad (18)$$

With

$$\Pr[f(\mathbf{X}) = 1 | \mathbf{X}_A = \mathbf{x}_A] = \frac{1}{2} + \frac{1}{2} \mathbb{E}[f(\mathbf{X}) | \mathbf{X}_A = \mathbf{x}_A]$$

and application of Lemma 1 given in Appendix A, (18) becomes

$$\begin{aligned} \sum_{S \subseteq A} \hat{f}(S) \Phi_S(\mathbf{x}_A) &= \hat{f}(\emptyset) \\ \Leftrightarrow \sum_{S \subseteq A \setminus \emptyset} \hat{f}(S) \Phi_S(\mathbf{x}_A) &= 0. \end{aligned} \quad (19)$$

It follows from the Fourier expansion (3), that (19) holds for all $\mathbf{x}_A \in \{-1, +1\}^{|A|}$ if and only if $\hat{f}(S) = 0$ for all $S \subseteq A \setminus \emptyset$, which proves the theorem.

E Proof of Theorem 4

First,

$$\begin{aligned} \Pr[f(\mathbf{X}) = 1 | \mathbf{X}_A = \mathbf{x}_A] &= \frac{1}{2} (1 + \mathbb{E}[f(\mathbf{X}) | \mathbf{X}_A = \mathbf{x}_A]) \\ &= \frac{1}{2} \underbrace{\left(1 + \sum_{S \subseteq A} \hat{f}(S) \Phi_S(\mathbf{x}_A) \right)}_{q(\mathbf{x}_A)}. \end{aligned} \quad (20)$$

where (20) follows from an application of Lemma 1. By definition of the conditional entropy,

$$\begin{aligned} H(f(\mathbf{X})|\mathbf{X}_A) &= \sum_{\mathbf{x}_A \in \{-1,1\}^{|A|}} \Pr[\mathbf{X}_A = \mathbf{x}_A] H(f(\mathbf{X})|\mathbf{X}_A = \mathbf{x}_A) \\ &= \sum_{\mathbf{x}_A \in \{-1,1\}^{|A|}} \Pr[\mathbf{X}_A = \mathbf{x}_A] h(\Pr[f(\mathbf{X}) = 1|\mathbf{X}_A = \mathbf{x}_A]) \\ &= \sum_{\mathbf{x}_A \in \{-1,1\}^{|A|}} \Pr[\mathbf{X}_A = \mathbf{x}_A] h(q(\mathbf{x}_A)) \end{aligned} \quad (21)$$

$$= \mathbb{E}[h(q(\mathbf{X}_A))] \quad (22)$$

where $h(\cdot)$ is the binary entropy function as defined in (5). To obtain (21), we used (20), and the expectation in (21) is with respect to the distribution of \mathbf{X}_A .

Observe that

$$\begin{aligned} \mathbb{E}[4q(\mathbf{X}_A)(1 - q(\mathbf{X}_A))] &= \mathbb{E}\left[1 - \left(\sum_{S \subseteq A} \hat{f}(S)\Phi_S(\mathbf{X}_A)\right)^2\right] \\ &= 1 - \sum_{S \subseteq A} \sum_{U \subseteq A} \hat{f}(S)\hat{f}(U)\mathbb{E}[\Phi_S(\mathbf{X}_A)\Phi_U(\mathbf{X}_A)] \\ &= 1 - \sum_{S \subseteq A} \hat{f}(S)^2, \end{aligned} \quad (23)$$

where (23) follows from the orthogonality of the basis functions.

We first prove the lower bound in Theorem 4. Applying the lower bound on the binary entropy function $h(p) \geq 4p(1 - p)$, given in [24, Thm. 1.2] on (22) yields

$$H(f(\mathbf{X})|\mathbf{X}_A) = \mathbb{E}[h(q(\mathbf{X}_A))] \geq \mathbb{E}[4q(\mathbf{X}_A)(1 - q(\mathbf{X}_A))],$$

and the lower bound in Theorem 4 follows using (23).

Next we prove the upper bound in Theorem 4. Applying the upper bound on the binary entropy function $h(p) \leq (p(1 - p))^{1/\ln(4)}$, given in [24, Thm. 1.2] on (22) yields

$$\begin{aligned} H(f(\mathbf{X})|\mathbf{X}_A) &= \mathbb{E}[h(q(\mathbf{X}_A))] \\ &\leq \mathbb{E}[\underbrace{(4q(\mathbf{X}_A)(1 - q(\mathbf{X}_A)))}_Y]^{1/\ln(4)}. \end{aligned} \quad (24)$$

The term Y in (24) is a random variable, and the function $(Y)^{1/\ln(4)}$ is concave in Y . An application of Jensen's inequality (see e.g. [19]) yields $\mathbb{E}[(Y)^{1/\ln(4)}] \leq (\mathbb{E}[Y])^{1/\ln(4)}$, hence the right-hand side of (24) can be lower as

$$H(f(\mathbf{X})|\mathbf{X}_A) \leq (\mathbb{E}[4q(\mathbf{X}_A)(1 - q(\mathbf{X}_A))])^{1/\ln(4)}. \quad (25)$$

Finally the upper bound in Theorem 4 follows from combining (25) and (23).

F Proof of Theorem 5

Equation (22) from the proof of Theorem 4 in Appendix E is

$$H(f(\mathbf{X})|\mathbf{X}_A) = \mathbb{E}[h(q(\mathbf{X}_A))]. \quad (26)$$

Inserting $q(\mathbf{X}_A)$ as given by (20) proves the Theorem.

References

- [1] S. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets, *J. Theor. Biol.* 22 (3) (1969) 437–467.
- [2] S. Kauffman, Homeostasis and differentiation in random genetic control networks, *Nature* 224 (5215) (1969) 177–178.
- [3] M. I. Davidich, S. Bornholdt, Boolean network model predicts cell cycle sequence of fission yeast, *PLoS ONE* 3 (2) (2008) e1672.
- [4] J. Saez-Rodriguez, L. Simeoni, J. A. Lindquist, R. Hemenway, U. Bommhardt, B. Arndt, U. Haus, R. Weismantel, E. D. Gilles, S. Klamt, B. Schraven, A logical model provides insights into t cell receptor signaling, *PLoS Comput. Biol.* 3 (8).
- [5] F. Li, T. Long, Y. Lu, Q. Ouyang, C. Tang, The yeast cell-cycle network is robustly designed, *Proc. Natl. Acad. Sci. USA* 101 (14) (2004) 4781–4786.
- [6] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, B. O. Palsson, Integrating high-throughput and computational data elucidates bacterial networks, *Nature* 429 (6987) (2004) 92–96.
- [7] J. Kesseli, P. Rämö, O. Yli-Harja, On spectral techniques in analysis of boolean networks, *Phys D Nonlinear Phenom* 206 (1-2) (2005) 49–61.
- [8] J. Kesseli, P. Rämö, O. Yli-Harja, Tracking perturbations in boolean networks with spectral methods, *Phys. Rev. E* 72 (2) (2005) 026137.
- [9] A. S. Ribeiro, S. A. Kauffman, J. Lloyd-Price, B. Samuelsson, J. E. S. Socolar, Mutual information in random boolean models of regulatory networks, *Phys. Rev. E* 77 (1) (2008) 011901.
- [10] G. Xiao, J. Massey, A spectral characterization of correlation-immune combining functions, *IEEE Trans. Inf. Theory* 34 (3) (1988) 569–571.
- [11] S. Liang, S. Fuhrman, R. Somogyi, Reveal, a general reverse engineering algorithm for inference of genetic network architectures, *Pac Symp Biocomput.* (1998) 18–29.
- [12] J. Grefenstette, S. Kim, S. Kauffman, An analysis of the class of gene regulatory functions implied by a biochemical model, *Biosystems*. 84 (2) (2006) 81–90.
- [13] B. Derrida, Y. Pomeau, Random networks of automata: A simple annealed approximation, *Europhys. Lett.* 1 (2) (1986) 49, 45.

- [14] R. R. Bahadur, A representation of the joint distribution of responses to n dichotomous items, in: *Studies in Item Analysis and Prediction*, Stanford University Press, (1961) 158–168.
- [15] M. Ben-Or, N. Linial, Collective coin flipping, robust voting schemes and minima of Banzhaf values, in: *Proc. of 26th IEEE Annu. Symp. on Found. of Comp. Science*, 1985, pp. 408–416.
- [16] N. H. Bshouty, C. Tamon, On the Fourier spectrum of monotone functions, *J. ACM* 43 (4) (1996) 747–770.
- [17] S. Schober, About boolean networks with noisy inputs, in: *Proc. of the 5th Int. Workshop on Comp. Syst. Biol. (WCSB 2008)*, 2008.
- [18] M. T. Matache, V. Matache, On the sensitivity to noise of a Boolean function, *J. Math. Phys.* 50 (10) (2009) 103512.
- [19] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, 2nd Edition, Wiley-Interscience, 2006.
- [20] B. Rosell, L. Hellerstein, S. Ray, Why skewing works: Learning difficult boolean functions with greedy tree learners, in: *Proc. of the 22nd Int. Conf. on Mach. Learn. 2005* (2005) 728–735.
- [21] A. Diskin, M. Koppel, Voting power: An information theory approach, *Soc. Choice and Welf.* 34 (1) (2010) 105–119.
- [22] R. Forre, Methods and instruments for designing S-boxes, *J. Cryptol.* 2 (3).
- [23] L. Raeymaekers, Dynamics of Boolean networks controlled by biologically meaningful functions, *J. Theor. Biol.* 218 (3) (2002) 331–341.
- [24] F. Topsøe, Bounds for entropy and divergence for distributions over a two-element set, *J. Ineq. Pure Appl. Math.* 2 (2001).